SFE2020 Evaluation Guide

GROUP A

A1. ISU ML Severe Wind Probs:
https://ousurvey.qualtrics.com/jfe/form/SV_3dYztPdne8Qi0tL
Research Questions:

1. Can machine-learning approaches provide useful information regarding the likelihood that wind-damage reports are associated with wind gusts >=50 knots?
2. Which machine-learning algorithms provide the most useful output? Why?
3. Are the ML severe wind probabilities often higher for measured gusts (despite measure gusts not being an input to the algorithms)?
4. Are the ML severe wind probabilities often higher for reports that occur in higher probabilistic wind forecasts from SPC Day 1 Outlooks (despite SPC outlooks not being an input to the algorithms)?
5. Are the ML severe wind probabilities often higher in more favorable environments (e.g., large instability and strong shear; which ***are*** inputs to the algorithms)?
6. Are the ML severe-wind probabilities often higher for more favorable convective modes (e.g., organized mesoscale convective systems; despite radar not being an input to the algorithms)?

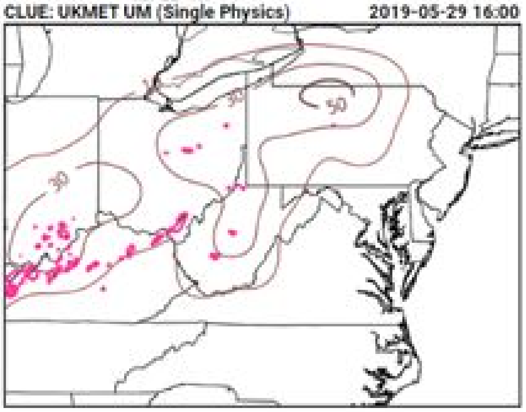
A2. NCAR ML Hazard Guidance: https://forms.gle/x8mKDvd2ZXA9gMp36
Research Questions:

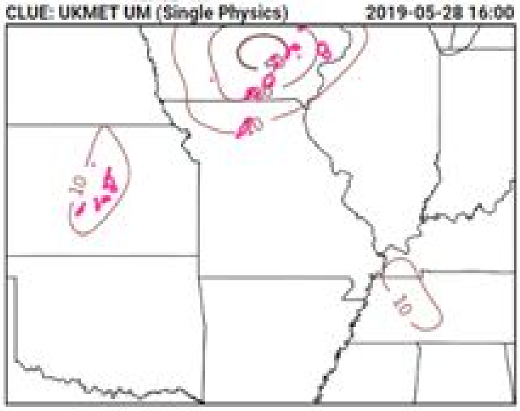
1. How well do the machine learning approaches (random forests and neural networks) compare to midlevel UH in probabilistic forecasts of severe weather?
2. Can machine-learning forecasts discriminate among severe weather hazards (i.e., tornado, hail, and wind)?
3. Which machine-learning approach, random forests (RF) or neural networks (NN), provided the most useful probabilistic guidance for total severe weather and the individual severe hazards: tornado, hail, and wind?
4. Do the machine-learning forecasts provide improved timing guidance over UH? If so, is there much difference between the RF and NN timing information?
5. Despite being derived from a deterministic forecast, would the machine-learning probabilities be a useful source of guidance to SPC forecasters (i.e., compare with 06Z SPC Day 1 Outlooks)?

A3. CLUE: 00Z CAM TL-Ensemble:
https://ousurvey.qualtrics.com/jfe/form/SV_8pk6sxxIQMXhEcR
Research Questions:

1. How does the performance of HREFv3 (replaces HRRRv3 with HRRRv4 and HRW NMMB with EMC FV3-SAR) compare to that of HREFv2.1?
2. How does the performance of the 00Z-initialized single-model ensembles (i.e., HRRRE and UM) compare to the HREF?
3. Do the single-model time-lagged ensembles (i.e., HRRRE TL-10 and UM TL-10) outperform their respective ensembles initialized at a single time (i.e., HRRRE and UM)?

4. How does the performance of the single-model time-lagged ensembles (i.e, HRRRE TL-10 and UM TL-10) compare to the HREF?
5. What is the relative importance of FAR and POD when assigning overall subjective performance ratings to CAM ensembles? (Note: POD answers the question “What is the fraction of events that occur where the probabilities are non-zero?” while FAR answers the question “Are the higher probabilities where they should be?”)





In the left image, FAR would likely be relatively more important than POD in assigning a lower rating while POD would likely be more important for assigning a higher rating in the right image.

A4. CLUE: TTU Ensemble Subsetting:
https://ousurvey.qualtrics.com/jfe/form/SV_5ngAZOlyipGHbBr
Research Questions:

1. How does the performance of the sensitivity-based HRRRE subset (6 members) compare to the full time-lagged HRRRE (9 members from 18Z and 9 members from 00Z = 18 members)?
2. How often are members initialized at 18Z selected for the ensemble subset?
3. Does the “best member” (i.e., member with lowest error in sensitive regions) add to or detract from the overall forecast guidance?
4. How often is one of the members initialized at 18Z selected as the “best member”?
5. Is the sensitivity-based ensemble subsetting approach a viable post-processing option in NWS operations to improve guidance for severe weather forecasting?

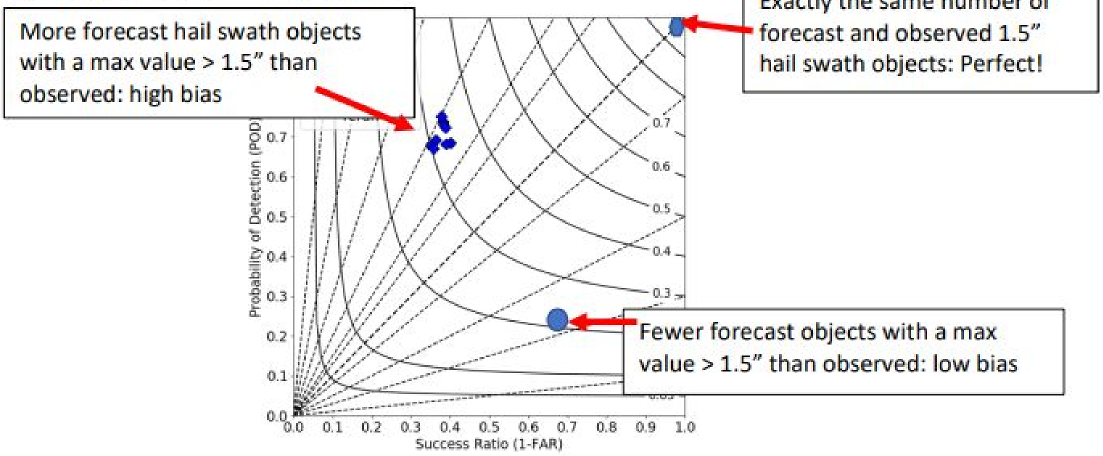
A5. CLUE: Ensemble Hail Guidance:
https://ousurvey.qualtrics.com/jfe/form/SV_aeFqv9kL9Le6Mvj
Research Questions:

1. What makes for a “good” hail forecast?
    a. Location accuracy
    b. Size accuracy
    c. Perspective: model developer vs. forecaster
    d. Temporal and spatial scale under consideration
2. Is the validation of hail forecasts over different time/spatial scales (e.g., 1-h “warning” scale, 6-h “watch” scale, 24-h “outlook” scale) helpful?
3. How well do these different scales capture hail forecast performance and how much do they differ?

4. How does the performance diagram with an object-based matching approach compare with a reliability diagram with a grid-based neighborhood approach? Do these different verification approaches provide different assessments of the "goodness" of the hail forecasts across different scales?
5. How does the performance of the different hail-size algorithms (HAILCAST, microphysics-based estimate) compare? Are there any systematic differences or biases that stand out for these algorithms?

Performance Diagram and Object-Matched Verification Method

This verification method uses MODE object-matching software to match observed and forecast swaths of hail. By using object-matching software, we can avoid penalizing the hail forecast if the underlying model failed to produce convection or put it in the wrong place. Once we have multiple matched hail swath objects, we can compare the maximum sizes from the forecast and observed swath objects. For example:



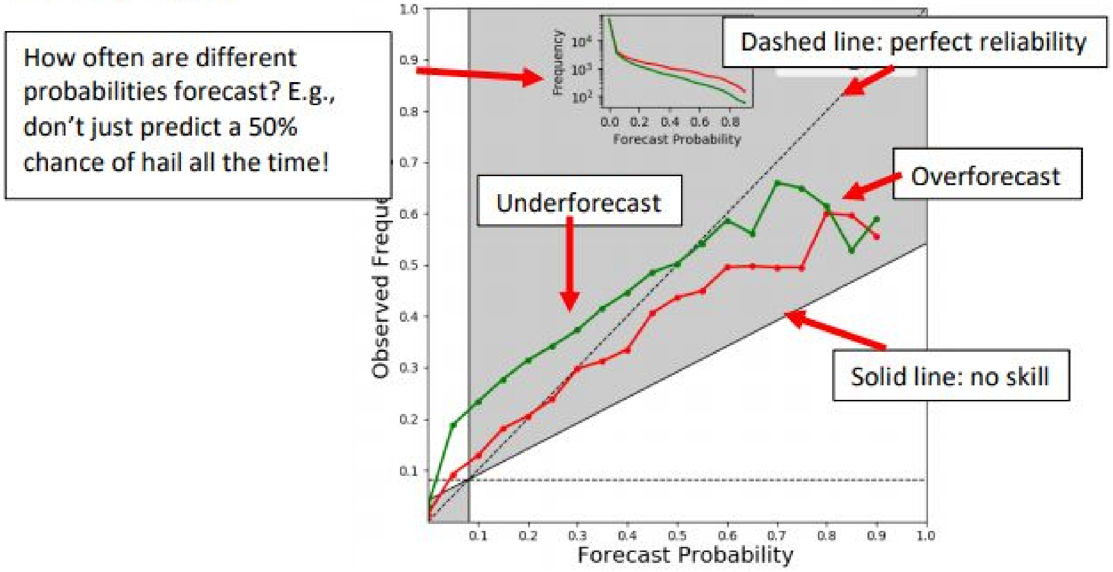
More forecast hail swath objects with a max value > 1.5" than observed: high bias

Exactly the same number of forecast and observed 1.5" hail swath objects: Perfect!

Fewer forecast objects with a max value > 1.5" than observed: low bias

Reliability Diagram and Grid-Based Verification Method

This method uses upscaling to account for convective displacement errors: a hail observation anywhere within a 40-km radius of a 1.5" hail forecast at a grid point is considered a "hit". The observed and forecast frequency of 1.5" hail at any grid point in the domain are then plotted on a reliability diagram:



How often are different probabilities forecast? E.g., don't just predict a 50% chance of hail all the time!

Dashed line: perfect reliability

Underforecast

Overforecast

Solid line: no skill

A6. CLUE: FV3-SAR Physics/DA/Vertical Levels:
https://ousurvey.qualtrics.com/jfe/form/SV_80qf8IQh01wfTh3
Research Questions:

1. What FV3-based configuration performs best during the first 12 h of forecasts?
   a. How long does the impact of radar DA last?
2. What FV3-based configuration produces the best depiction of severe convective storms?
   a. Timing
   b. Convective mode
   c. Location
   d. Storm depiction (i.e., storm size, intensity, number of storms, etc.) - here is likely where we'll see different responses for the appearance of of storms for the sarX memer with its updated physics (MYNN and Thompson)
3. What FV3-based configuration produces the best depiction of and evolution of storm environment fields?
   a. Instability
   b. Temperature
   c. Moisture
4. What differences in FV3-based CAMs lead to differences in the vertical sounding structure?
5. How does the performance of runs with different numbers of vertical levels compare (Note: NSSL FV3-SAR has 80 levels compared to 50 levels in EMC FV3-SARX)?

A7. CLUE: FV3-SAR IC/Hord/LSM (Continuation of A6 survey):
https://ousurvey.qualtrics.com/jfe/form/SV_80qf8IQh01wfTh3
Research Questions:

1. What FV3-based configuration produces the best depiction of severe convective storms?
   a. Timing
   b. Convective mode
   c. Location
   d. Storm depiction (i.e., storm size, intensity, number of storms, etc.)
2. What FV3-based configuration produces the best depiction of and evolution of storm environment fields?
   a. Instability
   b. Temperature
   c. Moisture
3. What impact does increased diffusivity have on model performance? (Note: hord=6 is more diffusive).
4. How does the performance of runs with different ICs/LBCs compare (Note: EMC using Noah and GSL using RUC LSM)?
5. What differences in FV3-based CAMs lead to differences in the vertical sounding structure?

A8. Mesoscale Analyses: https://ousurvey.qualtrics.com/jfe/form/SV_8J3XefsHHu6my8Z

Research Questions:

1. How useful are these analysis systems for situational awareness and assessment of the pre-convective and near-storm environment for convective weather applications?
2. Are there notable differences between the EMC and GSL versions of the 3D-RTMA?
    a. Which fields?
    b. When? Where? Why?
3. Does using information from a CAM ensemble in the hybrid-variational analysis improve its utility for short-term convective forecasting applications? (Note: EMC version gets ensemble background error covariance from the HRRRDAS while the GSL version uses the GDAS for BEC information.)
4. Are there any notable issues or artifacts in any of the fields?
    a. Ghosting or duplicate structures in the reflectivity field
    b. Circular bullseyes in surface fields, especially if not representative (i.e. poor QC)
    c. Geometric, irregular shapes in the CAPE fields

A9. Lightning DA: https://ousurvey.qualtrics.com/jfe/form/SV_ey9Z4CXc2O5R7wx

Research Questions:

1. Does the assimilation of GOES-16 GLM total lightning data improve the short-term forecasts of convection in areas with limited radar coverage?
2. If there is a positive impact of assimilating GLM data, approximately how long does it last into the forecast?

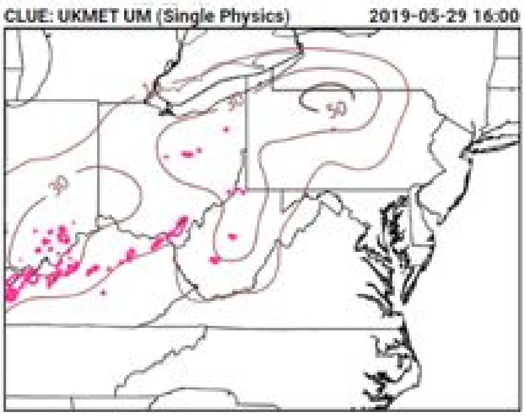
GROUP B

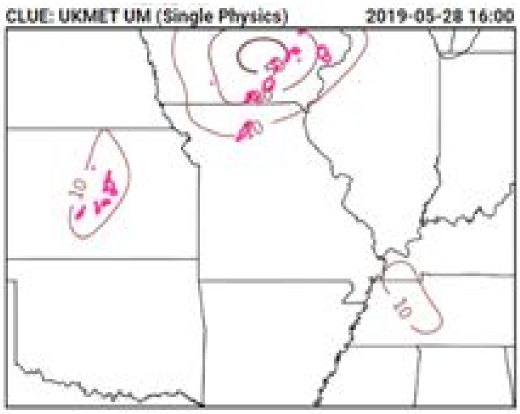
B1. HREF Calibrated Guidance:
https://ousurvey.qualtrics.com/jfe/form/SV_3yfRaQR00NdkjTT
Research Questions:

1. Which calibrated guidance method performs best? (Tor, Hail, Wind)
2. How does the SPC timing guidance perform relative to first-guess guidance? (Tor, Hail, Wind)
   a. What does the SPC timing guidance do better (worse) than the first-guess guidance?
3. How do the STP-calibrated tornado probabilities using STP values extracted from the inflow sector compare to the traditional STP-calibrated tornado probabilities (i.e., circular neighborhood)? (Tor)
4. How well does the MCS filter on the STP-calibrated tornado probabilities reduce probabilities in areas with a linear reflectivity mode? (Tor)
5. How appropriate are the magnitudes of the short-term STP-calibrated tornado probabilities? (Tor)
6. What differences are there between the ML methods for 24 h calibrated hail guidance? (Hail)
   a. Probability coverage, magnitude
7. For the ML products, does the Deep Learning method provide improved forecast guidance over the RF approaches? (Hail)
8. How do the ML hail methods compare to the HREF/SREF Calibrated guidance? (Hail)
9. How does ML wind guidance perform relative to HREF/SREF-calibrated guidance? (Wind)
   a. Probability coverage, magnitude

B2. CLUE: 00Z CAM Multi-Model Ensemble:
https://ousurvey.qualtrics.com/jfe/form/SV_86XKKvCoFgbol9P
Research Questions:

1. How does the performance of HREFv3 (replaces HRRRv3 with HRRRv4 and HRW NMMB with EMC FV3-SAR) compare to that of HREFv2.1?
2. How does the performance of a multi-model ensemble from a single initialization time compare to time-lagged single-model ensembles?
3. Can a time-lagging strategy for CAM ensemble design provide as much useful spread as a multi-model CAM configuration?
4. Does a time-lagged multi-model CAM ensemble provide the best probabilistic forecasts for severe weather applications?
5. Does the performance of the 36-member time-lagged multi-model ensemble (HRRRE+UM TL-36) meet or exceed that of the 10-member HREF?
6. What is the relative importance of FAR and POD when assigning overall subjective performance ratings to CAM ensembles? (Note: POD answers the question “What is the fraction of events that occur where the probabilities are non-zero?” while FAR answers the question “Are the higher probabilities where they should be?”)





In the left image, FAR would likely be relatively more important than POD in assigning a lower rating while POD would likely be more important for assigning a higher rating in the right image.

B3. CLUE: 12Z CAM TL-Ensemble:
https://ousurvey.qualtrics.com/jfe/form/SV_0dI1HvJltE6GknP
Research Questions:

1. Does the time-lagged ensemble (HRRRE TL-9) produce improved probabilistic forecasts (e.g., increased useful spread; less overconfidence) for severe weather applications over the traditional non-time-lagged ensemble (HRRRE)?
2. Does the multi-physics time-lagged ensemble (HRRR/NSSL WRF-TL) produce improved probabilistic forecasts for severe weather applications over the single-physics time-lagged ensemble? (Note: The HRRR/NSSL WRF-TL also has some additional IC diversity.)
3. Which of these ensembles produces the best probabilistic forecasts for severe weather applications?
4. Based on these results, what might be the best design strategy for a single-model-core CAM ensemble?

B4. Deterministic Flagships: **https://ousurvey.qualtrics.com/jfe/form/SV_3ZWgi1wkDrsF14h**
Research Questions:

1. How do current state-of-the-art deterministic CAMs storm attribute fields compare?
   a. Timing
   b. Convective mode
   c. Location
   d. Storm depiction (i.e., storm size, intensity, number of storms, etc.)
2. How do current state-of-the-art deterministic CAMs environmental fields compare?
   a. Instability
   b. Temperature
   c. Moisture
3. How does the relative performance of state-of-the-art deterministic CAMs change based on the time of day?
4. How well does the FV3 core compare to the WRF-ARW for convective weather applications?

B5. CLUE: Core and ICs (Continuation of B4 survey):
https://ousurvey.qualtrics.com/jfe/form/SV_3ZWgi1wkDrsF14h
Research Questions:

1. Does model core or ICs affect model forecasts of severe convective storms more?
2. Which set of ICs generates the best forecasts of:
   a. Storm attribute fields
   b. Environmental fields
3. Which model core generates the best forecasts of:
   a. Storm attribute fields
   b. Environmental fields
4. Does the impact of differing ICs wane after a certain period of time? If so, how long does that take?

B6. WoFS Configurations:
https://ousurvey.qualtrics.com/jfe/form/SV_aWbyeoDIWOwA1Cd
Research Questions:

A. How well is WoFS depicting severe convective storms?
   a. Feature timing
   b. Convective mode
B. How does WoFS performance change with different initialization cycles?
C. 3.0 km vs. 1.5 km system
   1. How does the experimental 1.5 km configuration of WoFS perform compared to the real-time 3 km system?
   2. What features do we see in the 1.5 km data that may not be obvious in the coarser data?
   3. Are there differences between the 1.5 km and 3.0 km data in CI timing or storm motion?
   4. Are the answers to any of the above questions a function of the initialization cycle?

D. Hybrid and Var WoFS Deterministic Runs
   1. Which deterministic WoFS configuration performs better?
   2. How do the deterministic runs compare to the RT ensemble configuration?
   3. Do the deterministic runs add value above the 3 km ensemble?

Short-Term Forecasting Evaluations
(NWS Forecasters only)

Evaluation of Yesterday's Forecasts: Innovation Desk:
https://ousurvey.qualtrics.com/jfe/form/SV_3lXElAvxbggbfMx
Research Questions:

1. How did the forecasts using WoFS perform relative to the forecasts without WoFS data?
2. What is the relative impact of WoFS across the different convective hazards?
3. Were 1-h forecasts or 4-h forecasts more skillful and why?
4. What effect did the updated WoFS guidance have on the forecasts?
5. What is the distribution of how much time it took to create these forecasts across a range of events?

WoFS/CAM Usage in Forecasting:
https://ousurvey.qualtrics.com/jfe/form/SV_bDwQgMrDUNDxzPT
Research Questions:

1. For which hazard was WoFS guidance most useful, and why?
2. Was there any particular hazard that was more difficult to forecast for than others?
3. Was there a difference in WoFS fields used to issue forecasts for differing hazards?
4. Were there any WoFS products that were used no matter the hazard type?
5. What data sources in addition to WoFS did forecasters use?
6. What factors increased or decreased confidence in the updated forecasts compared to the initial forecasts?
    a. WoFS? If so, what products?
    b. Observations? If so, what observations?
    c. A combination? E.g., CI occurred, so increased confidence in WoFS forecasts?

Evaluation of Yesterday's Forecasts: R2O Desk:
https://ousurvey.qualtrics.com/jfe/form/SV_2bphOw4vkSy4uYB
Research Questions:

1. How did the update forecasts using WoFS perform relative to the initial forecasts without WoFS data?
2. What was the relative impact of WoFS on the update forecasts across the different convective hazards?
3. How intuitive were the conditional intensity forecasts to draw?
4. What was the relative impact of WoFS on the conditional intensity forecasts across the different convective hazards?
5. Should operational SPC Outlooks replace significant severe coverage probabilities with conditional intensity bins?